From French EHR to NCI ontology via UMLS

Paolo Besana[1], Marc Cuggia[1], Oussama Zekri[2], Annabel Bourde[1],Anita Burgun[1]

[1] Université de Rennes 1, [2] Centre Régional de Lutte Contre Le Cancer - Eugène Marquis

Abstract. Clinical trials are required for evaluating new therapies and diagnostic techniques. We developed a system based on OWL and SWRL aimed at suggesting the clinical trials to which a patient could be enrolled, evaluating on real clinical trials and patients from the University Hospital of Rennes. This paper presents the method employed to map the French expressions from the patient data to terms in the NCI ontology.

1 Introduction

Clinical trials are fundamental for testing new therapies or diagnostic techniques. Patients are enrolled to a clinical trial if they match its eligibility criteria. We developed a work aimed at suggesting the clinical trials to which a patient could be enrolled [2]. We represent patient data from the Electronic Health Record (EHR) and the eligibility criteria using the NCI Thesaurus [4], an oncology-specific ontology developed by the National Institutes of Health. The project evaluation is based on the clinical trials and patients' data from the Centre Hospitalier Universitaire of Rennes (France) . This paper presents the work done to automatically map the French text of some of the EHR fields to terms from the NCI ontology.

2 Problem description

In order to evaluate the system we extracted 4 real clinical trials active during 2009 in the University Hospital of Rennes, and we selected 486 patients that were assessed for the trials during the same year. Both clinical trials and data were in French. The rules have been converted manually, while the size of the patients data required an automatic processing for at least some of the fields. This paper presents the method employed to automatically map the French expressions of some of the EHR fields to terms from the NCI ontology.

3 Method

A patient's record contains 44 relevant fields. We focussed on the fields containing single expressions. In particular, we chose the field specifying the site of the tumor, very relevant for the recruitment. We extracted the possible different source values $\mathbf{V_{src}} = \left\{v^1_{src}, ..., v^n_{src}\right\}$ of the field . All patients are admitted in urology, and all the cancers are related to urology: the possible values for the field are only 28. Of these values, 16 are composed by more than a single word (for example: 'col de l'uterus' meaning cervix). In particular 5 are specification of a position within an organ. We compared three methods for translation: MESH (Medical Subject Headings, terminology used to index PubMed publications) in French[1], Google translate and Wikipedia. In order to find the NCI concept,

[1] http://terminologiecismef.chu-rouen.fr/

Method	Translation found	Correct translation	CUIs found	NCI found	NCI correct
MESH	16 - 57%	16- 57%	13 - 46%	12 - 42%	12 - 42%
Wikipedia	20 - 71%	20 - 71%	21 - 75%	18 - 64%	15 - 53%
Google	28 - 100%	26 - 92%	14 - 50%	9 - 32%	7 - 25%

Table 1. Mapping results for the three different methods on 28 French terms. A translation is found if an expression is returned. The expressions are evaluated manually. The CUIs are concepts in UMLS found by string matching, and the NCI terms are linked to CUIs. The correctness of the NCI terms is checked manually.

we used the UMLS meta-thesaurus (Unified Medical Language System) [3]. In UMLS each concept is identified by a unique key (CUI). A concept is linked to many terms from different terminologies, including NCI. Each term in UMLS is defined by a semantic type (such as Disease or Body Part). Table 1 shows the results for the different methods, discussed more in detail below.

Using Google Translate Google translate was queried for each of the terms in $\mathbf{V_{src}}$. The only error in translation is for a term with homonyms in different domains. The translations were used to query the UMLS by string. To improve relevance, we filtered the results by semantic type. The terms specifying a position in an organ could not be matched directly to defined terms in any of the terminology contained in UMLS. The found CUIs were then filtered to extract the terms from NCI.

Using MESH French Each term in $\mathbf{V_{src}}$ is queried on the MESH French website. The result contains the corresponding English term from the original version of MESH. The English term is then used to query UMLS (filtering by string and by MESH terminology). Of the 16 found translations, 3 were terms belonging to different semantic types and therefore discarded. The resulting CUIs are queried to find the corresponding NCI term.

Using Wikipedia Inspired by the work in [1], each source term is queried in wikipedia French. A possible set of pages is returned. The pages are selected using their wikipedia categories. The filtered pages are used to extract the corresponding English page title. The mapping to NCI follows the same mechanism explained for Google translate. The use of categories to filter unrelated pages improves precision.

References

1. O. Collin B. Gaillard, M. Boualem. Query translation using wikipedia resources for analysis and disambiguation. In *Proceeding of EAMT 2010, Conference of the European Association for Machine Translation*, 2010.
2. P. Besana, M. Cuggia, O. Zekri, A. Bourde, and A. Burgun. Using semantic web technologies for clinical trial recruitment. In *International Semantic Web Conference 2010*, volume 6414. Springer, 2010.
3. O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database Issue):D267, 2004.
4. N.F. Noy, S. de Coronado, H. Solbrig, G. Fragoso, F.W. Hartel, and M.A. Musen. Representing the NCI Thesaurus in OWL DL: Modeling tools help modeling languages. *Applied ontology*, 3(3):173–190, 2008.